A REPORT

ON

Clustering Large Datasets with Gaussian Mixture Models for Traffic Analysis

Machine Learning (CSL7620) – Course Project

Submitted By:

Kachana Phanindra Reddy (M25IRM003)
Manish Kumar Milan (M25IRM005)
Somnath Bose (M25IRM008)
Group 24

Contents

1 Introduction **1**

2 Methods **1**

2.1 Data 1

2.1.1 Dataset Description: 1

2.1.2 Dataset Parsing: 1

2.2 Algorithm 2

2.2.1 Model Formulation: 2

2.2.2 Expectation-Maximization (EM) Algorithm: 2

2.2.3 Code Implementation: 3

3 Results **3**

Results **3**

3.1 Approach 1: Spatial-Based Clustering 3

3.2 Approach 2: Temporal-Based Clustering [2] 3

4 Conclusion **4**

1 INTRODUCTION

Traffic congestion significantly impacts mobility, energy efficiency, and safety, making the evaluation of traffic status a critical research focus. We try to use data-driven methods that offer computational efficiency and scalability for effective analysis of complex traffic dynamics without relying on restrictive theoretical assumptions.

This study employs traffic data from the **Performance Measurement System (PeMS)** developed by the **California Department of Transportation**. The raw data are preprocessed and modelled as two different setups: one for clustering sensors based on similar traffic history, another for clustering timesteps based on traffic states.

To classify traffic conditions, a **Gaussian Mixture Model (GMM)** is utilized due to its ability to model multimodal data distributions and capture the inherent variability of traffic states. Unlike conventional approaches, this fully data-driven, unsupervised framework enables interpretable traffic inference, which makes it suitable for traffic performance evaluation and intelligent transportation system applications.

2 METHODS

2.1 Data

2.1.1 Dataset Description:

The Performance Measurement System (PeMS) datasets are developed by the California Department of Transportation (Caltrans) and consist of real-time traffic data collected from loop detectors installed across freeways. In this study, two such datasets **PEMSD4** and **PEMSD8** are used for analysis.

Dataset	Location	# Sensors	Time Period	Frequency
PEMSD4	Bay Area	307	2 months	Every 5 minutes
PEMSD8	San Bernardino	170	2 months	Every 5 minutes

Each dataset contains three key traffic features:

- **Flow**: Number of vehicles passing the sensor within a given time interval.
- **Occupancy**: Percentage of time the sensor is covered by vehicles, serving as a congestion indicator.
- **Speed**: Average speed of vehicles over the sampling period.

The datasets are downloaded from `zenodo.org`[1] in `.npz` format, with data organized as a 3D tensor of dimensions `(timestep, sensor, feature)`.

2.1.2 Dataset Parsing:

Approach 1:

- The 3D dataset in 2D by concatenating data from all sensors into a single dataframe.
- The 5-minute sampling is then aggregated hourly for each day for every sensor, and then again aggregated by every hour of each day of the week for every sensor, averaging over the entire time period.

- Data transformed into a representation where every row refers to a unique sensor and contains its traffic feature history.

These feature vectors might be useful for clustering sensors with similar traffic patterns.

Approach 2:

- Only the dataframe for a single sensor is extracted.
- The 5-minute samples are then grouped hourly for each day by averaging.

These feature vectors might be useful for clustering sensors with similar traffic patterns.

2.2 Algorithm

A Gaussian Mixture Model (GMM) is a probabilistic clustering technique that assumes the data is generated from a mixture of several Gaussian (normal) distributions, each representing a distinct cluster.

2.2.1 Model Formulation:

A Gaussian mixture model with K components models the probability density function as:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

where the parameters $\lambda = \{\pi_k, \mu_k, \Sigma_k\}$ represent:

- π_k: Mixing coefficients (cluster weights)
- μ_k: Means of each Gaussian component
- Σ_k: Covariance matrices

The parameters are estimated by maximizing the data log-likelihood:

$$\log p(X|\lambda) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \right)$$

2.2.2 Expectation-Maximization (EM) Algorithm:

The EM algorithm is used to iteratively estimate these parameters:

E-step (Expectation): For each data point x_i and component k, compute the responsibility (posterior probability):

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}$$

M-step (Maximization): Given the responsibilities γ_{ik}, define $N_k = \sum_{i=1}^{N} \gamma_{ik}$ and update:

$$\pi_k = \frac{N_k}{N}, \quad \mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} x_i, \quad \Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^T$$

The process repeats until convergence of the log-likelihood.

2.2.3 Code Implementation:

- Data parsing is performed using `pandas`, and features are stored as NumPy arrays.
- Standardization is applied to all features.
- **Initialization:**
  - Mixing coefficients: $\pi_k = 1/K$
  - Means: from K-Means cluster centers (`init_method='kmeans'`) or random samples
  - Covariances: identity matrices or empirical covariances
- Dimensionality reduction is done using PCA, reducing 504 features to 45 (capturing 99% variance).
- Probabilities are computed using `scipy.stats.multivariate_normal.logpdf` for stability.
- Normalizing factors use `scipy.special.logsumexp`.
- Convergence is checked by monitoring the absolute change in log-likelihood between iterations.
- Visualizations are generated using `matplotlib` and `seaborn`.

3 RESULTS

3.1 Approach 1: Spatial-Based Clustering

In this spatial-based analysis, each sensor acts as an individual data point characterized by three primary traffic features like **flow**, **speed**, and **occupancy**, aggregated hourly across all days. After preprocessing and cleaning, the dataset contained 170 sensors with 504 features each. Since the feature dimension far exceeded the number of samples, **Principal Component Analysis (PCA)** was applied to reduce dimensionality while retaining 99% of the total variance, resulting in 45 principal components.

The **GMM** clustering algorithm was configured with three clusters. Each cluster corresponds to a set of sensors that have similar traffic patterns and histories over the period of study. This approach demonstrates that GMM combined with PCA effectively captures spatial variability across the sensor network. Sensors grouped within the same cluster exhibit similar traffic performance characteristics, enabling ba etter understanding of traffic dynamics across different roadway segments.

Visualization of cluster centroids in the reduced PCA space demonstrates well-separated Gaussian distributions, confirming that spatially neighboring sensors often belong to similar clusters. This implies that **GMM combined with PCA effectively captures spatial variability** across the sensor network. Furthermore, the model provides probabilistic membership scores for each sensor, allowing flexible interpretation of uncertain boundary cases (e.g., sensors transitioning between moderate and congested conditions).

3.2 Approach 2: Temporal-Based Clustering [2]

The Gaussian Mixture Model (GMM) clustering was applied on time-aggregated traffic data, where sensor readings of flow, speed, and occupancy were averaged hourly across 61 consecutive days. This temporal aggregation provided a consistent and representative view of daily traffic behavior over the study period. The

GMM algorithm was configured with two components, representing distinct traffic regimes observed throughout the day.

The results reveal that GMM effectively identifies two dominant traffic states, corresponding to normal (free-flow) and congested (crowded) conditions. The pairwise scatter plots — Flow vs Occupancy, Speed vs Occupancy, and Speed vs Flow — show well-separated clusters, illustrating the inherent nonlinear relationships among the three key features. The cluster centroids, indicated by black-edged markers, clearly highlight the mean characteristics of each traffic regime, providing interpretable distinctions between low-occupancy high-speed states and high-occupancy low-speed states.

Overall, this approach demonstrates that the GMM framework can successfully differentiate traffic states purely from observed measurements, enabling data-driven understanding of recurrent traffic dynamics. Combined with PCA-based spatial clustering, the methodology provides a unified view of both spatial and temporal traffic variability across the network.

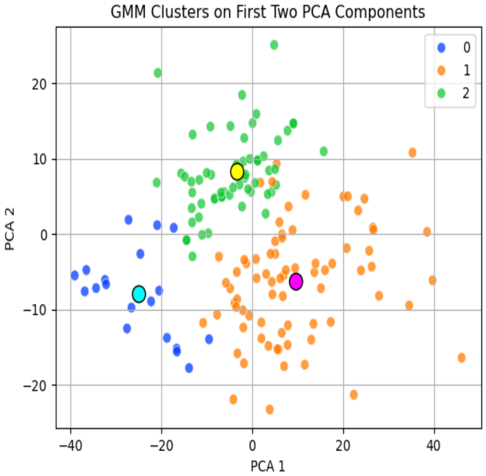

Figure 1: Approach 1

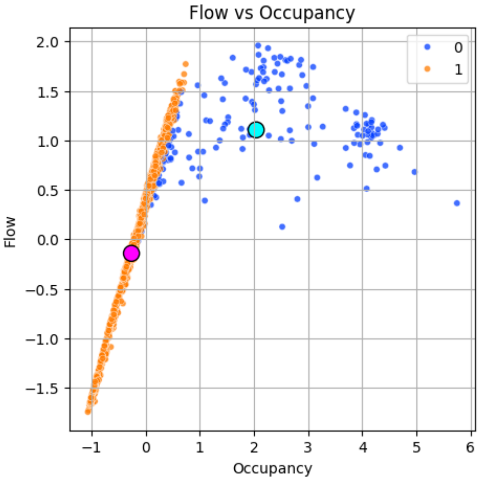

Figure 2: Approach 2

4 CONCLUSION

The Gaussian Mixture Model (GMM) effectively identified distinct traffic patterns from complex sensor data. Combined with PCA, it efficiently reduced dimensionality while preserving key variance. In the **spatial approach**, GMM grouped sensors with similar traffic behaviors, revealing regional flow similarities. In the **temporal approach**, it distinguished between normal and congested states based on flow, speed, and occupancy relationships. Overall, GMM with PCA provides a simple yet powerful framework for uncovering both spatial and temporal traffic patterns, aiding in better traffic analysis and management.

REFERENCES

[1] Shiqi Zhang, “TrafficDataSets,” *Zenodo*, Apr. 11, 2023. doi: 10.5281/zenodo.7816008.

[2] Xiong Liu, Li Pan, and Xiaoliang Sun. “Real-Time Traffic Status Classification Based on Gaussian Mixture Model.”